The Analysis of Two-Way Functional Data Using Two-Way Regularized Singular Value Decompositions

Jianhua Z. Huang, Haipeng Shen and Andreas Buja

Abstract

Two-way functional data consist of a data matrix whose row and column domains are both structured, for example, temporally or spatially, as when the data are time series collected at different locations in space. We extend one-way functional principal component analysis (PCA) to two-way functional data by introducing regularization of both left and right singular vectors in the singular value decomposition (SVD) of the data matrix. We focus on a penalization approach and solve the non-trivial problem of constructing proper two-way penalties from one-way regression penalties. We introduce conditional cross-validated smoothing parameter selection whereby left-singular vectors are cross-validated conditional on right-singular vectors, and vice versa. The concept can be realized as part of an alternating optimization algorithm. In addition to the penalization approach, we briefly consider two-way regularization with basis expansion. The proposed methods are illustrated with one simulated and two

Jianhua Z. Huang is Professor (Email: jianhua@stat.tamu.edu), Department of Statistics, Texas A&M University, College Station, TX 77843. Haipeng Shen (Email: haipeng@email.unc.edu) is Assistant Professor, Department of Statistics and Operations Research, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599. Andreas Buja (Email: buja@wharton.upenn.edu) is Liem Sioe Liong/First Pacific Company Professor, Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, PA 19104. Jianhua Z. Huang's work was partially supported by NSF grant DMS-0606580, NCI grant CA57030, and Award Number KUS-CI-016-04, made by King Abdullah University of Science and Technology (KAUST). Haipeng Shen's work was partially supported by NSF grant DMS-0606577, CMMI-0800575, and UNC-CH R. J. Reynolds Fund Award for Junior Faculty Development.

real data examples. Supplemental materials available online show that several "natural" approaches to penalized SVDs are flawed and explain why so.

Keywords: *Functional data analysis, penalization, regularization, spatial-temporal modelling, basis expansion*

1 Introduction

From a statistical modeling point of view, the main purpose of the SVD is to provide Least Squares (LS) fits of product terms or "rank-one approximations", $\lambda u_i v_j$, to suitably centered data matrices $x_{i,j}$. We can divide the uses of product terms into two major types, PCA and ANOVA, the first better known than the second, but both of interest to us:

- **PCA**: When $x_{i,j}$ is viewed as multivariate data (rows = iid multivariate samples), product terms are fitted to the column-centered data $x_{i,j} - m_j$. The values u_i are interpreted as one-dimensional projections of the cases or the estimates of a latent factor/predictor, while the values v_j are interpreted as forming the projection direction or the "loadings" of variable j on the "factor." The latent predictor interpretation stems from the "model" $x_{i,j} \approx m_j + \lambda u_i v_j$, where u_i plays the role of a shared predictor and v_j that of a column-specific slope.
- **ANOVA**: When $x_{i,j}$ is interpreted as a balanced two-way ANOVA table, the SVD can be used to fit product interactions (Williams, 1952; Mandel, 1971). The product term is fitted to the residuals of an additive fit, $x_{i,j} - m - a_i - b_j \approx \lambda u_i v_j$. If further analysis reveals $u_i \approx f(a_i)$ and $v_j \approx g(b_j)$, one has found a non-linear Tukey one-degree of freedom interaction of the form $\lambda f(a_i) g(b_j)$.

Extending the first of these two uses, the SVD has also become an important tool in functional data analysis (FDA, Ramsay and Silverman (2002, 2005)) where each row of the data matrix is thought of as discretized values of a function evaluated at

some common grid points, one function per row. The grid points are elements of a continuous domain such as space or time. Because the domain is continuous, one assumes the functions that generate the rows to be smooth, and the goal of FDA is to incorporate such assumptions. Even before the advent of FDA it was common to apply plain PCA to functional data, early examples being Rao (1958, 1987).

To impose smoothness on PCA coefficients, Rice and Silverman (1991) and Silverman (1996) introduced regularization through roughness penalties on the eigenvectors. Of the two approaches, Silverman (1996)'s is more principled and works as follows: Given a data matrix $\mathbf{X} = (x_{i,j})_{i\in I, j\in J}$, one assumes that the domain J is structured, usually as space or time, with an implied notion of smoothness. The degree of smoothness of coefficient vectors $\mathbf{v} = (v_j)_{j\in J}$ can be measured by quadratic penalties $\mathbf{v}^T\mathbf{\Omega}\mathbf{v}$ ($\mathbf{\Omega} = (\Omega_{j',j''})_{j',j''\in J}$) that may be as simple as sums of squared second differences, $\mathbf{v}^T\mathbf{\Omega}\mathbf{v} = \sum(2v_j - v_{j-1} - v_{j+1})^2$, in case of a time domain. Assuming the columns of $\mathbf{X}$ centered as needed, plain PCA maximizes the Rayleigh quotient $\mathcal{R}_0(\mathbf{v}) = \|\mathbf{X}\mathbf{v}\|^2/\|\mathbf{v}\|^2$. Silverman (1996) proposes regularization by penalizing the denominator as follows: $\mathcal{R}(\mathbf{v}) = \|\mathbf{X}\mathbf{v}\|^2/(\|\mathbf{v}\|^2 + \mathbf{v}^T\mathbf{\Omega}\mathbf{v})$. Hence he maximizes variance with regard to a regularized reference norm that inflates for non-smooth vectors, thus favoring smooth vectors for large eigenvalues. Our proposed method specializes to that of Silverman (1996), and we provide a reformulation of functional PCA using SVD in Huang et al. (2008).

In this article, we deal with data that are functional in two ways: In $\mathbf{X} = (x_{i,j})_{i\in I, j\in J}$ both index domains I and J are structured with notions of smoothness. We thereby leave the strict domain of multivariate analysis where rows are considered iid samples, and we move closer to a two-way ANOVA interpretation of the data, but with a non-standard model in mind. This can be motivated with the two real data examples considered in Section 6:

- Section 6.2 deals with a demographic application where the data matrix records

mortality rates for different age groups in the United States from 1959 to 1999. It is reasonable to assume that the mortality rate is a smooth function of both age and time period.

- Section 6.3 is concerned with the "patience" or willingness to wait of customers who made calls to a telephone call center. Customer patience is measured using the logit transformation of the survival function of time-willing-to wait, which we expect to depend smoothly on both time of day and waiting time.

Both examples exhibit a two-way functional structure where neither way represents iid samples, thus resembling more an ANOVA situation. To describe the two-way functional structure, we view the element x_{ij} of the data matrix $\mathbf{X}$ as evaluation of an underlying function $X(\cdot, \cdot)$ on a rectangular grid of sampling points (y_i, z_j), where y_i $(i = 1, \ldots, n)$ are from a domain $\mathcal{Y}$ and z_j $(j = 1, \ldots, m)$ are from a domain $\mathcal{Z}$. Because we require a symmetric treatment of the domains, we cannot rely on PCA and its asymmetric treatment of rows and columns in its eigendecomposition. We are therefore led to the SVD which offers symmetric treatment. More specifically, we use the fact that the SVD provides low-rank approximations to the data matrix.

To approximate the two-way functional data using r components of product terms, a continuous version of a partial SVD model is as follows:

$$X(y, z) = U_1(y)V_1(z) + U_2(y)V_2(z) + \cdots + U_r(y)V_r(z) + \epsilon(y, z)\,, \tag{1}$$

where we absorbed the singular value λ_k into $U_k(y)$ and/or $V_k(z)$, and where the error is iid white noise. We assume that $U_k(y)$ and $V_k(z)$ are smooth on their respective domains, and it is this two-fold smoothness requirement that we incorporate in two-way regularized SVDs. As written, the model should be interpreted as a functional fixed-effects model where the functions are fixed but unknown. — We consider primarily regularization with roughness penalties, but we will also discuss regularization with basis expansion (Section 4), all of which we refer to as "regularized SVDs."

It is possible to interpret one or both components in each product term of (1) as random instead of fixed effects and apply time series or spatial models, in particular when time or space dependence is more naturally interpreted in terms of auto-correlation. For example, Hyndmann and Booth (2008) considered V_k's as fixed, smooth functions and U_k's as random effects subject to time series modeling. Apparently, the fixed-effects/smoothing and the random-effects/time series views provide two different modeling frameworks for the same kind of data and each has its own merit. We shall focus on the fixed effects/smoothing view in this paper. An exception is the Bayes model outlined in Section 2.4, which can be interpreted as a hierarchical model whereby nature draws discretized functions from the prior and adds iid noise before presenting them to the observer.

Regularization with penalization is widely used in statistics (Wahba, 1990; Green and Silverman, 1994) and in machine learning ("kernelizing"; Schölkopf and Smola (2001)). Nevertheless, the application of two-way penalization to SVDs is not a trivial matter. Section 2.2 derives the proper form of penalization from axiomatic conditions. Further topics are the reduction of 2-way penalized SVD to an ordinary SVD with "half-smoothing" (Section 2.3), Bayes priors (Section 2.4), and Reproducing Kernel Hilbert Space (RKHS) theory that connects penalization on finite-dimensional data spaces and on function spaces (Section 2.5). The latter two sections are brief but they point to potentially far-reaching generalizations with Bayes approaches and kernelizing techniques. Next we approach smoothing parameter selection for the two penalties in terms of left-right conditional cross-validation (Section 3.1). Conditioning on left and right singular vectors alternatingly spares us the need to estimate two smoothing parameters simultaneously. Left-right conditional cross-validation can be justified as leave-one-out operations on the rows and columns of $\mathbf{X}$ (Section 3.2). Section 4 discusses the basis expansion approach. Section 5 derives a formal equivalence between two-way penalized SVDs and penalized canonical correlations (Leurgans et al.,

1993) using the notion of a "bi-Rayleigh quotient" that generalizes squared canonical correlations. Section 6 presents a simulation and the two real data examples.

2 The Structure of Penalized SVDs

Our discussion focuses on extracting the first pair of components in (1); subsequent pairs can be extracted sequentially by removing the effect of preceding pairs.

2.1 Unpenalized LS for rank-one approximation

We write rank-one approximations to a $n \times m$ data matrix $\mathbf{X}$ as $\mathbf{u}\mathbf{v}^T$, where $\mathbf{u}$ and $\mathbf{v}$ are n- and m-vectors, respectively. We will not assume that either is normalized, hence they are determined only up to a scale factor that can be shifted between them:

$$\mathbf{u} \mapsto c\mathbf{u}\,, \quad \mathbf{v} \mapsto \mathbf{v}/c \quad (c \neq 0). \tag{2}$$

Writing $\|\mathbf{M}\|^2 = \sum_{i,j} M_{i,j}^2$ for the squared Frobenius norm of an arbitrary matrix $\mathbf{M}$, the unregularized LS criterion for rank-one approximations is

$$\mathcal{C}_0(\mathbf{u}, \mathbf{v}) \;=\; \|\mathbf{X} - \mathbf{u}\mathbf{v}^T\|^2 \;=\; \|\mathbf{X}\|^2 - 2\mathbf{u}^T\mathbf{X}\mathbf{v} + \|\mathbf{u}\|^2\|\mathbf{v}\|^2. \tag{3}$$

The problem can be cast as two conditional sets of LS problems whose solutions are

$$\text{argmin}_{\mathbf{u}}\, \mathcal{C}_0(\mathbf{u}, \mathbf{v}) \;=\; \frac{\mathbf{X}\mathbf{v}}{\|\mathbf{v}\|^2} \quad \text{and} \quad \text{argmin}_{\mathbf{v}}\, \mathcal{C}_0(\mathbf{u}, \mathbf{v}) \;=\; \frac{\mathbf{X}^T\mathbf{u}}{\|\mathbf{u}\|^2}\,. \tag{4}$$

They express the fact that, for fixed $\mathbf{v}$, the optimal $\mathbf{u}$ consists of the set of slopes of simple linear regressions (without intercept) of each row of $\mathbf{X}$ onto $\mathbf{v}$ (the shared single predictor); similarly, for fixed $\mathbf{u}$, the optimal $\mathbf{v}$ results from regressing each column of $\mathbf{X}$ onto $\mathbf{u}$. These equations can be used to justify the power algorithm

$$\mathbf{u} \leftarrow \mathbf{X}\mathbf{v}\,, \quad \mathbf{v} \leftarrow \mathbf{X}^T\mathbf{u}\,, \quad \text{followed by normalizations,} \tag{5}$$

which — if initialized randomly — converges almost surely to a LS rank-one fit.

2.2 Penalized LS for rank-one approximation

We introduce domain-specific penalty matrices $\mathbf{\Omega}_u$ $(n \times n)$ and $\mathbf{\Omega}_v$ $(m \times m)$, both symmetric and non-negative definite, whose purpose is to balance goodness-of-fit as measured by $\mathcal{C}_0(\mathbf{u}, \mathbf{v})$ against smoothness as measured by the penalties $\mathbf{u}^T\mathbf{\Omega}_u\mathbf{u}$ and $\mathbf{v}^T\mathbf{\Omega}_v\mathbf{v}$. Penalty matrices are usually endowed with multipliers α_u and α_v, the smoothing parameters (also referred to as penalty parameters or bandwidths for short); for now we absorb them into $\mathbf{\Omega}_u$ and $\mathbf{\Omega}_v$ and defer their selection with cross-validation to Section 3. Associated with the penalties are smoother matrices

$$\mathbf{S}_u = (\mathbf{I} + \mathbf{\Omega}_u)^{-1}, \quad \mathbf{S}_v = (\mathbf{I} + \mathbf{\Omega}_v)^{-1}$$

(Hastie and Tibshirani, 1990) which solve, respectively,

$$\mathbf{S}_u\mathbf{y} \;=\; \text{argmin}_{\mathbf{u}} \left(\|\mathbf{y} \,-\, \mathbf{u}\|^2 + \mathbf{u}^T\mathbf{\Omega}_u\mathbf{u} \right), \qquad \mathbf{S}_v\mathbf{z} \;=\; \text{argmin}_{\mathbf{v}} \left(\|\mathbf{z} \,-\, \mathbf{v}\|^2 + \mathbf{v}^T\mathbf{\Omega}_v\mathbf{v} \right).$$

We now pose the problem of finding a penalized criterion for rank-one approximation:

$$\mathcal{C}(\mathbf{u}, \mathbf{v}) \;=\; \|\mathbf{X} \,-\, \mathbf{u}\mathbf{v}^T\|^2 \;+\; \mathcal{P}(\mathbf{u}, \mathbf{v}), \tag{6}$$

where the penalty $\mathcal{P}(\mathbf{u}, \mathbf{v})$ is to be determined. A requirement we impose is that the minimizing $\mathbf{u}$ and $\mathbf{v}$ are conditionally smoothed versions of the LS solutions (4):

$$\text{argmin}_{\mathbf{u}}\, \mathcal{C}(\mathbf{u}, \mathbf{v}) \;\propto\; \mathbf{S}_u\mathbf{X}\mathbf{v} \quad \text{and} \quad \text{argmin}_{\mathbf{v}}\, \mathcal{C}(\mathbf{u}, \mathbf{v}) \;\propto\; \mathbf{S}_v\mathbf{X}^T\mathbf{u}\,. \tag{7}$$

These conditions interlock the smoothing of $\mathbf{u}$ and $\mathbf{v}$, as becomes clear by considering the alternating power algorithm similar to (5) which they justify:

$$\mathbf{u} \leftarrow \mathbf{S}_u\mathbf{X}\mathbf{v}\;, \quad \mathbf{v} \leftarrow \mathbf{S}_v\mathbf{X}^T\mathbf{u}\;, \quad \text{followed by normalizations.} \tag{8}$$

One may ask about this algorithm a) whether it converges, and, if so, b) whether it minimizes any criterion at all, and, if so, c) whether this criterion amounts to a form of penalized least squares of the form (6). All these questions can be answered in the affirmative, but the solution is not obvious. The following theorem (proof in the appendix) uniquely characterizes the only two-way penalty $\mathcal{P}(\mathbf{u}, \mathbf{v})$ that can be said to simultaneously penalize $\mathbf{u}$ according to $\mathbf{\Omega}_u$ and $\mathbf{v}$ according to $\mathbf{\Omega}_v$:

Theorem 1 *Assume $\mathcal{P}(\mathbf{u}, \mathbf{v})$ has the following properties:*

(i) $\mathbf{u} \mapsto \mathcal{P}(\mathbf{u}, \mathbf{v})$ *is a quadratic for fixed* $\mathbf{v}$*, and* $\operatorname{argmin}_{\mathbf{u}} \mathcal{C}(\mathbf{u}, \mathbf{v}) \propto \mathbf{S}_u \mathbf{X} \mathbf{v}$.

(ii) $\mathbf{v} \mapsto \mathcal{P}(\mathbf{u}, \mathbf{v})$ *is a quadratic for fixed* $\mathbf{u}$*, and* $\operatorname{argmin}_{\mathbf{v}} \mathcal{C}(\mathbf{u}, \mathbf{v}) \propto \mathbf{S}_v \mathbf{X}^T \mathbf{u}$.

(iii) *If* $\mathbf{\Omega}_u = \mathbf{0}$ *and* $\mathbf{\Omega}_v = \mathbf{0}$*, then* $\mathcal{P} \equiv 0$.

Then $\mathcal{P}(\mathbf{u}, \mathbf{v})$ has the following form:

$$\mathcal{P}(\mathbf{u}, \mathbf{v}) \;=\; \mathbf{u}^T \mathbf{\Omega}_u \mathbf{u} \cdot \|\mathbf{v}\|^2 \;+\; \|\mathbf{u}\|^2 \cdot \mathbf{v}^T \mathbf{\Omega}_v \mathbf{v} \;+\; \mathbf{u}^T \mathbf{\Omega}_u \mathbf{u} \cdot \mathbf{v}^T \mathbf{\Omega}_v \mathbf{v}$$

For future reference we write the penalty and the criterion in the following forms:

$$\mathcal{P}(\mathbf{u}, \mathbf{v}) \;=\; \mathbf{u}^T (\mathbf{I} + \mathbf{\Omega}_u) \mathbf{u} \cdot \mathbf{v}^T (\mathbf{I} + \mathbf{\Omega}_v)\, \mathbf{v} - \|\mathbf{u}\|^2 \|\mathbf{v}\|^2 \tag{9}$$

$$\mathcal{C}(\mathbf{u}, \mathbf{v}) \;=\; \|\mathbf{X} - \mathbf{u}\mathbf{v}^T\|^2 \;+\; \mathbf{u}^T \mathbf{\Omega}_u \mathbf{u}\, \|\mathbf{v}\|^2 \;+\; \|\mathbf{u}\|^2\, \mathbf{v}^T \mathbf{\Omega}_v \mathbf{v} \;+\; \mathbf{u}^T \mathbf{\Omega}_u \mathbf{u}\, \mathbf{v}^T \mathbf{\Omega}_v \mathbf{v} \tag{10}$$

$$\;=\; \|\mathbf{X}\|^2 \;-\; 2\mathbf{u}^T \mathbf{X} \mathbf{v} \;+\; \mathbf{u}^T (\mathbf{I} + \mathbf{\Omega}_u) \mathbf{u} \cdot \mathbf{v}^T (\mathbf{I} + \mathbf{\Omega}_v)\, \mathbf{v} \tag{11}$$

From (11) we obtain the exact stationary equations for $\mathcal{C}(\mathbf{u}, \mathbf{v})$ for later use:

$$\operatorname{argmin}_{\mathbf{u}} \mathcal{C}(\mathbf{u}, \mathbf{v}) = \frac{\mathbf{S}_u \mathbf{X} \mathbf{v}}{\mathbf{v}^T (\mathbf{I} + \mathbf{\Omega}_v) \mathbf{v}}\,, \qquad \operatorname{argmin}_{\mathbf{v}} \mathcal{C}(\mathbf{u}, \mathbf{v}) = \frac{\mathbf{S}_v \mathbf{X}^T \mathbf{u}}{\mathbf{u}^T (\mathbf{I} + \mathbf{\Omega}_u) \mathbf{u}}\,. \tag{12}$$

The criterion $\mathcal{C}(\mathbf{u}, \mathbf{v})$ has some desirable properties: (i) Scale invariance under (2): $\mathcal{C}(c\mathbf{u}, \mathbf{v}) = \mathcal{C}(\mathbf{u}, c\mathbf{v})$; (ii) Equivariance under rescaling of $\mathbf{X}$ and the fit $\mathbf{u}\mathbf{v}^T$: $\mathcal{C}(c\mathbf{u}, \mathbf{v}; c\mathbf{X}) = \mathcal{C}(\mathbf{u}, c\mathbf{v}; c\mathbf{X}) = c^2\, \mathcal{C}(\mathbf{u}, \mathbf{v}; \mathbf{X})$; (iii) For $\mathbf{\Omega}_u = 0$, the penalty specializes to the one-way penalty of Silverman (1996); (iv) The stationary equations of $\mathcal{C}(\mathbf{u}, \mathbf{v})$ involve smoothing with penalties $\mathbf{\Omega}_u$ and $\mathbf{\Omega}_v$, not scalar multiples thereof. Several "natural" approaches to penalizing SVDs do not share some of these properties, see the supplementary materials. We show these flawed approaches to spare readers fruitless search in dead ends.

2.3 Penalized SVDs are generalized SVDs via half-smoothing

The penalized SVD based on $\mathcal{C}(\mathbf{u}, \mathbf{v})$ is a plain SVD in a non-standard coordinate system. The new coordinates $\tilde{\mathbf{u}}$ and $\tilde{\mathbf{v}}$ are linked to the original coordinates $\mathbf{u}$ and $\mathbf{v}$

in terms of the "half-smoothers" $\mathbf{S}_u^{1/2} = (\mathbf{I} + \mathbf{\Omega}_u)^{-1/2}$ and $\mathbf{S}_v^{1/2} = (\mathbf{I} + \mathbf{\Omega}_v)^{-1/2}$:

$$\mathbf{S}_u^{1/2}\,\tilde{\mathbf{u}} \;=\; \mathbf{u} \quad \text{and} \quad \mathbf{S}_v^{1/2}\,\tilde{\mathbf{v}} \;=\; \mathbf{v}\,. \tag{13}$$

Let $\tilde{\mathbf{X}} = \mathbf{S}_u^{1/2}\,\mathbf{X}\,\mathbf{S}_v^{1/2}$ be the data matrix $\mathbf{X}$ half-smoothed two-ways in rows and columns. The penalized SVD criterion (11) can then be rewritten as

$$\begin{aligned} \mathcal{C}(\mathbf{u},\mathbf{v}) &= \|\mathbf{X}\|^2 \;-\; 2\tilde{\mathbf{u}}^T\tilde{\mathbf{X}}\tilde{\mathbf{v}} \;+\; \|\tilde{\mathbf{u}}\|^2 \;\cdot\; \|\tilde{\mathbf{v}}\|^2 \\ &= \|\mathbf{X}\|^2 \;-\; \|\tilde{\mathbf{X}}\|^2 \;+\; \|\tilde{\mathbf{X}} - \tilde{\mathbf{u}}\tilde{\mathbf{v}}^T\|^2, \end{aligned}$$

which is equivalent to the unpenalized LS criterion (3) for the plain SVD on the transformed matrix $\tilde{\mathbf{X}}$. This extends Silverman (1996)'s observation from one-way to two-way regularized SVDs. Thus there is something to the intuition that the data matrix $\mathbf{X}$ can be smoothed directly, but the proper steps are

1. to half-smooth the data matrix according to $\tilde{\mathbf{X}} = \mathbf{S}_u^{1/2}\,\mathbf{X}\,\mathbf{S}_v^{1/2}$,
2. to obtain a plain SVD of the half-smoothed data matrix $\tilde{\mathbf{X}}$, and
3. to half-smooth the singular vectors according to $\mathbf{S}_u^{1/2}\tilde{\mathbf{u}} = \mathbf{u}$ and $\mathbf{S}_v^{1/2}\tilde{\mathbf{v}} = \mathbf{v}$.

As the penalized SVD is an ordinary SVD in non-standard coordinates, the notions of orthogonality and length are non-standard under penalization. While for $\tilde{\mathbf{u}}$ and $\tilde{\mathbf{v}}$ the inner products and squared norms are Euclidean, for $\mathbf{u}$ and $\mathbf{v}$ they are:

$$\begin{aligned} \langle\!\langle \mathbf{u}_1, \mathbf{u}_2 \rangle\!\rangle \;&=\; \tilde{\mathbf{u}}_1^T\,\tilde{\mathbf{u}}_2 \;=\; \mathbf{u}_1^T\,(\mathbf{I} + \mathbf{\Omega}_u)\,\mathbf{u}_2\,, \qquad [\![\mathbf{u}]\!]^2 \;=\; \tilde{\mathbf{u}}^T\,\tilde{\mathbf{u}} \;=\; \mathbf{u}^T\,(\mathbf{I} + \mathbf{\Omega}_u)\,\mathbf{u}\,, \\ \langle\!\langle \mathbf{v}_1, \mathbf{v}_2 \rangle\!\rangle \;&=\; \tilde{\mathbf{v}}_1^T\,\tilde{\mathbf{v}}_2 \;=\; \mathbf{v}_1^T\,(\mathbf{I} + \mathbf{\Omega}_v)\,\mathbf{v}_2\,, \qquad [\![\mathbf{v}]\!]^2 \;=\; \tilde{\mathbf{v}}^T\,\tilde{\mathbf{v}} \;=\; \mathbf{v}^T\,(\mathbf{I} + \mathbf{\Omega}_v)\,\mathbf{v}\,, \end{aligned}$$

extending another of Silverman (1996)'s observations to the two-way case.

The iterative power algorithm (5) is not generally used for calculating the ordinary SVD. The above discussion suggests application of efficient SVD algorithms (Golub and van Loan, 1996) for calculating our penalized SVD. However, the conditional view in the power algorithm is critical for us to identify an appropriate penalized criterion (Theorem 1, Section 2.2) and to develop a cross-validation criterion for smoothing parameter selection for the penalized SVD (Section 3).

2.4 Bayes priors for rank-one approximation

The Bayes interpretation of penalized smoothing stems from the normal likelihood $\mathbf{y}|\mathbf{f} \sim \mathcal{N}(\mathbf{f}, \sigma^2\mathbf{I})$ and the prior $\mathbf{f} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{\Sigma})$. With $\mathbf{\Omega} = \mathbf{\Sigma}^{-1}$ and $\mathbf{S} = (\mathbf{I} + \mathbf{\Omega})^{-1}$, the posterior is $\mathbf{f}|\mathbf{y} \sim \mathcal{N}(\mathbf{Sy}, \sigma^2\mathbf{S})$. In the Bayes view the penalty matrix is (up to a scale factor) the inverse of the prior covariance, and eigendirections of $\mathbf{\Omega}$ with zero eigenvalues are interpreted as carrying an improper flat prior. The smooth $\mathbf{Sy}$ is the conditional posterior mean of $\mathbf{f}$ given the data $\mathbf{y}$.

The two-way penalty for rank-one approximations proposed here calls for interpretation as a Bayes prior, but the prior for $(\mathbf{u}, \mathbf{v})$ implied by (9),

$$p(\mathbf{u}, \mathbf{v}) \;\propto\; \exp\left(-\frac{1}{2\sigma^2}\left(\mathbf{u}^T(\mathbf{I} + \mathbf{\Omega}_u)\mathbf{u} \cdot \mathbf{v}^T(\mathbf{I} + \mathbf{\Omega}_v)\mathbf{v} - \|\mathbf{u}\|^2\|\mathbf{v}\|^2\right)\right),$$

is improper for two reasons: First, it is flat for $(\mathbf{u}, \mathbf{v})$ that have $\mathbf{\Omega}_u\mathbf{u} = \mathbf{0}$ and $\mathbf{\Omega}_v\mathbf{v} = \mathbf{0}$, as it should be. Second, its form is analogous to $\exp(-\frac{1}{2}x^2y^2)$ which is integrable in x for $y \neq 0$ and vice versa, but not jointly integrable. Underlying this second reason is invariance under the scale transformation (2): $\mathbf{u} \mapsto c\mathbf{u},\ \mathbf{v} \mapsto \mathbf{v}/c$, which the above density satisfies. This improper prior produces the desired partial posteriors,

$$\begin{aligned}
\mathbf{u}\,|\,\mathbf{X}, \mathbf{v} &\;\sim\; \mathcal{N}\left(\frac{1}{\mathbf{v}^T\mathbf{S}_v\mathbf{v}}\mathbf{S}_u\mathbf{X}\mathbf{v}\,,\ \frac{1}{\mathbf{v}^T\mathbf{S}_v\mathbf{v}}\mathbf{S}_u\right), \\
\mathbf{v}\,|\,\mathbf{X}, \mathbf{u} &\;\sim\; \mathcal{N}\left(\frac{1}{\mathbf{u}^T\mathbf{S}_u\mathbf{u}}\mathbf{S}_v\mathbf{X}^T\mathbf{u}\,,\ \frac{1}{\mathbf{u}^T\mathbf{S}_u\mathbf{u}}\mathbf{S}_v\right),
\end{aligned}$$

which invite Gibbs sampling as an alternative to alternating conditional mean estimation with smoothing. Gibbs sampling will be more interesting for models that involve distributions other than normal, such as general exponential families.

2.5 Functional SVD and RKHS Theory

So far we treated the penalized SVD problem in finite dimensions. For a truly functional view we need to connect the penalty introduced above to a penalty on function spaces. The standard framework for this purpose is Reproducing Kernel Hilbert

Space (RKHS) theory. We start by assuming $\mathbf{X} = (X(y_i, z_j))_{i=1,\ldots,n;\, j=1,\ldots,m}$ to contain the evaluations of a realization of a random field $X(y, z)$ at (y_i, z_j), where y_i and z_j are distinct sampling points in the respective domains $\mathcal{Y}$ and $\mathcal{Z}$. We seek a functional rank-one or product approximation $X(y, z) \simeq U(y)\, V(z)$. We assume the index domains $\mathcal{Y}$ and $\mathcal{Z}$ of $U(y)$ and $V(z)$ are endowed with RKHSs $\mathcal{H}_u$ and $\mathcal{H}_v$ to which $U(y)$ and $V(z)$ are confined. The RKHSs carry reproducing kernels $K_u(y_1, y_2)$ and $K_v(z_1, z_2)$, inner products $\langle U_1, U_2 \rangle_u$ and $\langle V_1, V_2 \rangle_v$, as well as norms $\|U\|_u$ and $\|V\|_v$, respectively. A fundamental property of RKHSs is that evaluations $U \mapsto U(y)$ are continuous functionals that can be represented by the kernel as follows: $\langle K_u(y, \cdot), U(\cdot) \rangle_u = U(y)$, and similarly for $V(z)$. According to the usual representer argument of Kimeldorf and Wahba (1971), there exists for arbitrary $\mathbf{u} = (u_1, \ldots, u_n)^T \in \mathbb{R}^n$ a unique $U \in \mathcal{H}_u$ with $u_i = U(y_i)$ $(i = 1, \ldots, n)$ and minimum norm $\|U\|_u$. Furthermore, this function is of the form $U(y) = \sum_{i=1,\ldots,n} c_i K_u(y_i, y)$, and $\|U\|_u^2 = \mathbf{u}^T \mathbf{\Omega}_u \mathbf{u}$, where $\mathbf{\Omega}_u = \mathbf{K}_u^{-1}$ and $\mathbf{K}_u = (K_u(y_{i'}, y_{i''}))_{i', i''=1,\ldots,n}$. (An identical argument yields $V(z) = \sum_{j=1,\ldots,m} d_j K_v(z_j, z)$ for given $\mathbf{v} \in \mathbb{R}^m$, and $\mathbf{\Omega}_v = \mathbf{K}_v^{-1}$.) The translation of the criterion $\mathcal{C}(\mathbf{u}, \mathbf{v})$ (10) to RKHSs is (by abuse of notation):

$$\begin{aligned} \mathcal{C}(U, V) \;&=\; \|\mathbf{X} - \mathbf{u}\mathbf{v}^T\|^2 \\ &\quad +\; \|U\|_u^2\, \|\mathbf{v}\|^2 \;+\; \|\mathbf{u}\|^2\, \|V\|_v^2 \;+\; \|U\|_u^2\, \|V\|_v^2\,, \end{aligned} \tag{14}$$

where $\mathbf{u} = (U(y_1), \ldots, U(y_n))^T$ and $\mathbf{v} = (V(z_1), \ldots, V(z_m))^T$. The representer argument shows that the finite-dimensional minimizers $\mathbf{u}$ and $\mathbf{v}$ of $\mathcal{C}(\mathbf{u}, \mathbf{v})$ translate to RKHS minimizers U and V of $\mathcal{C}(U, V)$. For special cases, such as cubic smoothing splines on finite intervals with $\|U\|_u^2 = \int \{U(y)''\}^2 dy$, see standard references such as Green and Silverman (1994).

3 Cross-validation

3.1 Conditional bandwidth selection with GCV

So far we have dealt with the "fixed bandwidth" problem, that is, fixed penalty matrices. We now discuss adaptive bandwidth selection for the criterion $\mathcal{C}(\mathbf{u}, \mathbf{v})$. We make bandwidths explicit as α_u and α_v in $\mathcal{C}(\mathbf{u}, \mathbf{v})$,

$$\begin{aligned} \mathcal{C}(\mathbf{u}, \mathbf{v}; \alpha_u, \alpha_v) \;&=\; \|\mathbf{X} - \mathbf{u}\mathbf{v}^T\|^2 \;+\; \mathbf{u}^T(\alpha_u \mathbf{\Omega}_u)\mathbf{u} \cdot \|\mathbf{v}\|^2 \\ &\quad +\; \|\mathbf{u}\|^2 \cdot \mathbf{v}^T(\alpha_v \mathbf{\Omega}_v)\mathbf{v} \;+\; \mathbf{u}^T(\alpha_u \mathbf{\Omega}_u)\mathbf{u} \cdot \mathbf{v}^T(\alpha_v \mathbf{\Omega}_v)\mathbf{v}\;, \end{aligned} \tag{15}$$

but we will drop the arguments α_u and α_v if the context allows. We will denote the smoothers associated with $\alpha_u \mathbf{\Omega}_u$ and $\alpha_v \mathbf{\Omega}_v$ by, respectively,

$$\mathbf{S}_u(\alpha_u) \;=\; (\mathbf{I} + \alpha_u \mathbf{\Omega}_u)^{-1}, \qquad \mathbf{S}_v(\alpha_v) \;=\; (\mathbf{I} + \alpha_v \mathbf{\Omega}_v)^{-1}.$$

Among methods for adaptive bandwidth choice we will focus on generalized cross-validation (GCV) and, for heuristics, on leave-one-out cross-validation (LOO-CV). [For a discussion, see for example Hastie and Tibshirani (1990), Section 3.4.] For a linear smoother $\mathbf{S}(\alpha)$ such that $\hat{\mathbf{y}} = \mathbf{S}(\alpha)\mathbf{y}$, the GCV score is defined as

$$\mathrm{GCV}(\alpha) \;=\; \frac{\frac{1}{n}\,\|\hat{\mathbf{y}} - \mathbf{y}\|^2}{\{1 - \frac{1}{n}\,\mathrm{tr}\,\mathbf{S}(\alpha)\}^2} \;=\; \frac{\frac{1}{n}\,\|\{\mathbf{I} - \mathbf{S}(\alpha)\}\mathbf{y}\|^2}{\{1 - \frac{1}{n}\,\mathrm{tr}\,\mathbf{S}(\alpha)\}^2}. \tag{16}$$

We discuss below how to define the GCV in our setting by making a connection of penalized SVD to linear smoothers.

To avoid simultaneous minimization of two bandwidth parameters, we nest bandwidth selection inside the alternating algorithm that optimizes $\mathbf{u}$ for fixed $\mathbf{v}$ and $\mathbf{v}$ for fixed $\mathbf{u}$. The steps involve smoothing with $\mathbf{S}_u(\alpha_u)$ and $\mathbf{S}_v(\alpha_v)$, respectively, with adaptively selected bandwidths α_u and α_v.

A point to keep in mind is that all procedures, updates as well as bandwidth selections, should be kept invariant under scale changes (2). The penalized LS criterion $\mathcal{C}(\mathbf{u}, \mathbf{v})$ will be minimized by a uniquely scaled rank-one matrix $\mathbf{u}\mathbf{v}^T$, but the directions $\mathbf{u}$ and $\mathbf{v}$ will be identifiable only up to a factor: $(c\mathbf{u})(\mathbf{v}/c)^T = \mathbf{u}\mathbf{v}^T$. Thus

alternating minimization will converge to a correctly sized solution $\mathbf{u}\mathbf{v}^T$ where the relative sizes of the two factors depend on the initialization. The alternating updates, with proper relative scale, are obtained from the stationary equations (12):

$$\mathbf{u} \;=\; \frac{\mathbf{S}_u(\alpha_u)\,\mathbf{X}\mathbf{v}}{\mathbf{v}^T(\mathbf{I}+\alpha_v\boldsymbol{\Omega}_v)\mathbf{v}} \;=\; \frac{\mathbf{S}_u(\alpha_u)}{1+\alpha_v\mathcal{R}_v(\mathbf{v})}\,\frac{\mathbf{X}\mathbf{v}}{\|\mathbf{v}\|^2}\,, \tag{17}$$

$$\mathbf{v} \;=\; \frac{\mathbf{S}_v(\alpha_v)\,\mathbf{X}^T\mathbf{u}}{\mathbf{u}^T(\mathbf{I}+\alpha_u\boldsymbol{\Omega}_u)\mathbf{u}} \;=\; \frac{\mathbf{S}_v(\alpha_v)}{1+\alpha_u\mathcal{R}_u(\mathbf{u})}\,\frac{\mathbf{X}^T\mathbf{u}}{\|\mathbf{u}\|^2}\,, \tag{18}$$

where $\mathcal{R}_u(\mathbf{u}) = \mathbf{u}^T\boldsymbol{\Omega}_u\mathbf{u}/\|\mathbf{u}\|^2$ and $\mathcal{R}_v(\mathbf{v}) = \mathbf{v}^T\boldsymbol{\Omega}_v\mathbf{v}/\|\mathbf{v}\|^2$ are the plain Rayleigh quotients of $\boldsymbol{\Omega}_u$ and $\boldsymbol{\Omega}_v$, and $\alpha_u\mathcal{R}_u(\mathbf{u})$ and $\alpha_v\mathcal{R}_v(\mathbf{v})$ those of $\alpha_u\boldsymbol{\Omega}_u$ and $\alpha_v\boldsymbol{\Omega}_v$, respectively. A comparison of (17) and (18) with the updates (5) of the unpenalized LS problem,

$$\mathbf{u} \;=\; \frac{\mathbf{X}\mathbf{v}}{\|\mathbf{v}\|^2}\,, \qquad \mathbf{v} \;=\; \frac{\mathbf{X}^T\mathbf{u}}{\|\mathbf{u}\|^2},$$

shows that the action in each iteration is not just smoothing but also shrinking that depends on the amount of penalization of the input component. Thus it is $\mathbf{S}_u(\alpha_u)/(1+\alpha_v\mathcal{R}_v(\mathbf{v}))$ and $\mathbf{S}_v(\alpha_v)/(1+\alpha_u\mathcal{R}_u(\mathbf{u}))$, not $\mathbf{S}_u(\alpha_u)$ and $\mathbf{S}_v(\alpha_v)$, that must be used when forming cross-validation criteria similar to (16) for selecting α_u and α_v:

$$\mathrm{GCV}_u(\alpha_u;\alpha_v) \;=\; \frac{\frac{1}{n}\,\Big\|\,\Big\{\mathbf{I}-\frac{\mathbf{S}_u(\alpha_u)}{1+\alpha_v\mathcal{R}_v(\mathbf{v})}\Big\}\frac{\mathbf{X}\mathbf{v}}{\|\mathbf{v}\|^2}\,\Big\|^2}{\Big\{1-\frac{1}{n}\,\mathrm{tr}\,\frac{\mathbf{S}_u(\alpha_u)}{1+\alpha_v\mathcal{R}_v(\mathbf{v})}\Big\}^2}\,, \tag{19}$$

$$\mathrm{GCV}_v(\alpha_v;\alpha_u) \;=\; \frac{\frac{1}{m}\,\Big\|\,\Big\{\mathbf{I}-\frac{\mathbf{S}_v(\alpha_v)}{1+\alpha_u\mathcal{R}_u(\mathbf{u})}\Big\}\frac{\mathbf{X}^T\mathbf{u}}{\|\mathbf{u}\|^2}\,\Big\|^2}{\Big\{1-\frac{1}{m}\,\mathrm{tr}\,\frac{\mathbf{S}_v(\alpha_v)}{1+\alpha_u\mathcal{R}_u(\mathbf{u})}\Big\}^2}\,. \tag{20}$$

If, according to the update formulas (17) and (18), we set $\mathbf{u} = \mathbf{S}_u(\alpha_u)\mathbf{X}\mathbf{v}/\mathbf{v}^T(\mathbf{I}+\alpha_v\boldsymbol{\Omega}_v)\mathbf{v}$ in (19) and $\mathbf{v} = \mathbf{S}_v(\alpha_v)\mathbf{X}^T\mathbf{u}/\mathbf{u}^T(\mathbf{I}+\alpha_u\boldsymbol{\Omega}_u)\mathbf{u}$ in (20), where $\mathbf{v}$ is seen as input and $\mathbf{u}$ as output in the first equation, and vice versa in the second, we get

$$\mathrm{GCV}_u(\alpha_u;\alpha_v) = \frac{\frac{1}{n}\,\Big\|\,\frac{\mathbf{X}\mathbf{v}}{\|\mathbf{v}\|^2}-\mathbf{u}\,\Big\|^2}{\Big\{1-\frac{1}{n}\,\frac{\mathrm{tr}\,\mathbf{S}_u(\alpha_u)}{1+\alpha_v\mathcal{R}_v(\mathbf{v})}\Big\}^2}\,, \quad \mathrm{GCV}_v(\alpha_v;\alpha_u) = \frac{\frac{1}{m}\,\Big\|\,\frac{\mathbf{X}^T\mathbf{u}}{\|\mathbf{u}\|^2}-\mathbf{v}\,\Big\|^2}{\Big\{1-\frac{1}{m}\,\frac{\mathrm{tr}\,\mathbf{S}_v(\alpha_v)}{1+\alpha_u\mathcal{R}_u(\mathbf{u})}\Big\}^2}\,. \tag{21}$$

A curious aspect of the nested approach to bandwidth selection is that the GCV scores are based on residuals not from the data matrix $\mathbf{X}$, but from its projections $\mathbf{X}\mathbf{v}$ and $\mathbf{X}^T\mathbf{u}$. Such indirect and conditional bandwidth selection may not be implausible

because a vector $\mathbf{v}$ that is smooth on its index domain may still result in a projection $\mathbf{Xv}$ that is rough on the index domain of $\mathbf{u}$, and vice versa; smoothness on both index domains needs to be enforced separately. Yet, it is natural to ask whether there exists a GCV method that works off the data matrix $\mathbf{X}$. In the next subsection, we work out cross-validation by deleting a column or row from the original data matrix and, to our surprise, we arrive at the same GCV scores given above. The derivation provides a formal justification of (21) that has so far been obtained heuristically.

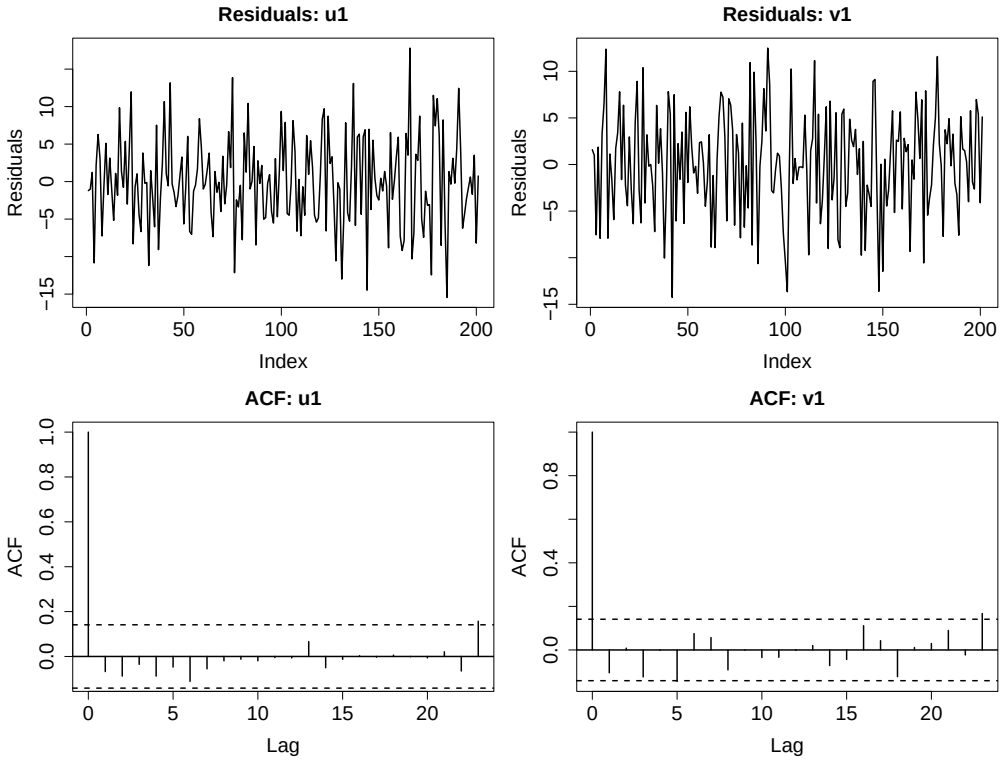


Figure 1: Residuals and corresponding ACFs for the two conditional regressions in extracting the first pair of components in a two product term model.

Remark. It is well known that cross-validation does not work well in smoothing problems with correlated errors, as the omitted point is not independent of the data used to fit the model. There is no correlated error problem in our fixed effect model, however. The residuals in the conditional regressions underlying the numerators

of GCVs (21), $\mathbf{X}\mathbf{v}/\|\mathbf{v}\|^2 - \mathbf{u}$ and $\mathbf{X}^T\mathbf{u}/\|\mathbf{u}\|^2 - \mathbf{v}$, do not have auto-correlations. This is true even when there are multiple product terms in (1). As an illustration, Figure 1 shows the residuals and the corresponding auto-correlation functions (ACF) for the two conditional regressions in extracting the first pair of components for a data set simulated from a two product term model of Section 6.1. To gain further understanding, note that (1) is a fixed effects model with white noise errors. When there is only one product term, absence of error correlation follows from the model assumption. When there are multiple product terms, because of orthogonality within the U components and within the V components, the conditional regressions with the U or V components fixed have approximate orthogonal design matrices and therefore boil down to a collection of simple regressions with uncorrelated errors.

3.2 Derivation of GCV from leaving out rows and columns

The criterion $\mathcal{C}(\mathbf{u}, \mathbf{v})$, when holding $\mathbf{u}$ fixed, can be viewed as an LS criterion with a generalized ridge penalty, where $\mathbf{u}$ determines a predictor matrix $\bar{\mathbf{X}}$, the data matrix $\mathbf{X}$ is strung out to form a response vector $\bar{\mathbf{y}}$, the vector $\mathbf{v}$ contains the regression coefficients, and the penalty $\mathcal{P}(\mathbf{u}, \mathbf{v})$ determines the ridge penalty matrix $\mathbf{\Omega}_{v|u}$:

$$\bar{\mathbf{y}} \;=\; \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \cdots \\ \mathbf{x}_m \end{pmatrix}, \qquad \bar{\mathbf{X}} \;=\; \begin{pmatrix} \mathbf{u} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{u} & \cdots & \mathbf{0} \\ \cdots & \cdots & \cdots & \cdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{u} \end{pmatrix},$$

$$\mathbf{\Omega}_{v|u} \;=\; (\mathbf{u}^T\mathbf{\Omega}_u\mathbf{u})\,\mathbf{I} + \|\mathbf{u}\|^2\,\mathbf{\Omega}_v + (\mathbf{u}^T\mathbf{\Omega}_u\mathbf{u})\,\mathbf{\Omega}_v\,,$$

where $\mathbf{x}_j$ is the j-th column of $\mathbf{X}$, and where $\bar{\mathbf{y}}$ is of size $mn \times 1$ and $\bar{\mathbf{X}}$ is of size $mn \times m$. Both the design matrix $\bar{\mathbf{X}}$ and the ridge penalty depend on $\mathbf{u}$. It is immediate that

$$\mathcal{C}(\mathbf{u}, \mathbf{v}) \;=\; \|\,\bar{\mathbf{y}} - \bar{\mathbf{X}}\mathbf{v}\,\|^2 \;+\; \mathbf{v}^T\,\mathbf{\Omega}_{v|u}\,\mathbf{v}\,, \tag{22}$$

which is a penalized LS criterion for $\mathbf{v}$. The associated penalized covariance is

$$\bar{\mathbf{X}}^T\bar{\mathbf{X}} + \mathbf{\Omega}_{v|u} \;=\; (\mathbf{u}^T(\mathbf{I} + \mathbf{\Omega}_u)\,\mathbf{u})\,(\mathbf{I} + \mathbf{\Omega}_v)\,,$$

and its inverse is

$$(\bar{\mathbf{X}}^T\bar{\mathbf{X}} + \boldsymbol{\Omega}_{v|u})^{-1} \;=\; \frac{1}{\mathbf{u}^T(\mathbf{I}+\boldsymbol{\Omega}_u)\,\mathbf{u}}\,\mathbf{S}_v\,.$$

Thus the hat matrix of the ridge regression is

$$\mathbf{H} \;=\; \bar{\mathbf{X}}\,(\bar{\mathbf{X}}^T\bar{\mathbf{X}} + \boldsymbol{\Omega}_{v|u})^{-1}\,\bar{\mathbf{X}}^T \;=\; \frac{1}{\mathbf{u}^T(\mathbf{I}+\boldsymbol{\Omega}_u)\,\mathbf{u}}\,\bar{\mathbf{X}}\,\mathbf{S}_v\,\bar{\mathbf{X}}^T.$$

Consider now cross-validation that leaves out one column of $\mathbf{X}$ at a time. Let $\hat{\mathbf{v}} = (\hat{v}_1, \ldots, \hat{v}_m)^T$ be the $\mathbf{v}$ that minimizes (22), and $\hat{\mathbf{v}}^{(-j)} = (\hat{v}_1^{(-j)}, \ldots, \hat{v}_m^{(-j)})^T$ be the same when the j-th block of $\bar{\mathbf{y}}$ and the corresponding rows of $\bar{\mathbf{X}}$ are removed. Then:

Lemma 1 *The j-th leave-out-one-column cross-validated error sum of squares is*

$$\|\mathbf{u}\hat{v}_j^{(-j)} - \mathbf{x}_j\|^2 \;=\; \mathbf{x}_j^T\mathbf{x}_j - \frac{(\mathbf{x}_j^T\mathbf{u})^2}{\|\mathbf{u}\|^2} + \|\mathbf{u}\|^2\,\frac{\left(\hat{v}_j - \mathbf{u}^T\mathbf{x}_j/\|\mathbf{u}\|^2\right)^2}{\{1-(\mathbf{S}_v)_{jj}/(1+\mathcal{R}_u(\mathbf{u}))\}^2}. \qquad (23)$$

Because we are holding $\mathbf{u}$ fixed, the first two terms in (23) are irrelevant. Averaging the last term in (23) over j and ignoring the factor $\|\mathbf{u}\|^2$, and re-introducing the weighted penalties $\alpha_u\boldsymbol{\Omega}_u$ and $\alpha_v\boldsymbol{\Omega}_v$, the cross-validation criterion for α_v is as follows:

$$\mathrm{CV}_v(\alpha_v;\alpha_u) \;=\; \frac{1}{m}\sum_{j=1}^{m}\frac{(\hat{v}_j - \mathbf{x}_j^T\mathbf{u}/\|\mathbf{u}\|^2)^2}{\{1-[\mathbf{S}_v(\alpha_v)]_{jj}/(1+\alpha_u\mathcal{R}_u(\mathbf{u}))\}^2}\,.$$

Replacing $[\mathbf{S}_v(\alpha_v)]_{jj}$ by its average over j, which is $\operatorname{tr}\mathbf{S}_v(\alpha_v)/m$, we obtain the generalized cross-validation criterion:

$$\mathrm{GCV}_v(\alpha_v;\alpha_u) \;=\; \frac{\frac{1}{m}\left\|\hat{\mathbf{v}} - \frac{\mathbf{X}^T\mathbf{u}}{\|\mathbf{u}\|^2}\right\|^2}{\left\{1-\frac{1}{m}\,\frac{\operatorname{tr}\mathbf{S}_v(\alpha_v)}{1+\alpha_u\mathcal{R}_u(\mathbf{u})}\right\}^2}\,.$$

A dual result holds for $\mathrm{CV}_u(\alpha_u;\alpha_v)$, which completes the derivation of (21).

4 Basis expansion

Crellin (1996) considered regularized SVD through basis expansions by minimizing

$$\|\mathbf{X} - \mathbf{u}\mathbf{v}^T\|^2 \qquad \text{subject to } \;\mathbf{u} = \mathbf{B}_u\boldsymbol{\phi},\; \mathbf{v} = \mathbf{B}_v\boldsymbol{\psi}, \qquad (24)$$

where $\mathbf{B}_u = (\mathbf{b}_{u1}, \ldots, \mathbf{b}_{uk})$ $(n\times k)$, $\mathbf{B}_v = (\mathbf{b}_{v1}, \ldots, \mathbf{b}_{vl})$ $(m\times l)$, $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_k)^T$ $(k\times 1)$ and $\boldsymbol{\psi} = (\psi_1, \ldots, \psi_l)^T$ $(l\times 1)$, Regularization is achieved by restricting $\mathbf{u}$ and $\mathbf{v}$ to

low dimensional subspaces $\mathcal{V}_u = \text{span}(\mathbf{b}_{u1}, \mathbf{b}_{u2}, ..., \mathbf{b}_{uk})$ and $\mathcal{V}_v = \text{span}(\mathbf{b}_{v1}, \mathbf{b}_{v2}, ..., \mathbf{b}_{vl})$, respectively. Subspace restrictions can be interpreted as limiting cases of penalizations by using squared distances from the subspaces $\mathcal{V}_u$ and $\mathcal{V}_v$ as penalties: $\alpha_u \|(\mathbf{I}-\mathbf{P}_u)\mathbf{u}\|^2$ and $\alpha_v \|(\mathbf{I}-\mathbf{P}_v)\mathbf{v}\|^2$, where $\mathbf{P}_u = \mathbf{B}_u(\mathbf{B}_u^T\mathbf{B}_u)^{-1}\mathbf{B}_u^T$ and $\mathbf{P}_v = \mathbf{B}_v(\mathbf{B}_v^T\mathbf{B}_v)^{-1}\mathbf{B}_v^T$ are the orthogonal projections onto $\mathcal{V}_u$ and $\mathcal{V}_v$, respectively. Up to scalar factors, the associated penalty matrices are the residual operators, $\mathbf{\Omega}_u = \alpha_u (\mathbf{I} - \mathbf{P}_u)$ and $\mathbf{\Omega}_v = \alpha_v(\mathbf{I} - \mathbf{P}_v)$, respectively, where $\alpha_u,\ \alpha_v \to \infty$. Less stringent penalty parameters $\alpha_u < \infty$ and $\alpha_v < \infty$ can be used to shrink the solution toward these subspaces.

Computations for minimizing (24) are trivial if one orthonormalizes the bases of $\mathcal{V}_u$ and $\mathcal{V}_v$. By abuse of notation we denote the orthonormalized bases again $\mathbf{b}_{u1}, ..., \mathbf{b}_{uk}$ and $\mathbf{b}_{v1}, ..., \mathbf{b}_{vl}$, so that the matrices satisfy $\mathbf{B}_u^T\mathbf{B}_u = \mathbf{I}$ and $\mathbf{B}_v^T\mathbf{B}_v = \mathbf{I}$, and the orthogonal projections are $\mathbf{P}_u = \mathbf{B}_u\mathbf{B}_u^T$ and $\mathbf{P}_v = \mathbf{B}_v\mathbf{B}_v^T$. With these prerequisites, the problem (24) boils down to a plain SVD on the projected data $\tilde{\mathbf{X}} = \mathbf{B}_u^T\mathbf{X}\mathbf{B}_v$ $(k \times l)$:

$$\mathcal{C}(\boldsymbol{\phi}, \boldsymbol{\psi}) \ = \ \|\mathbf{X}\|^2 \ - \ 2\,\boldsymbol{\phi}^T\mathbf{B}_u^T\mathbf{X}\mathbf{B}_v\boldsymbol{\psi} \ + \ \|\boldsymbol{\phi}\|^2\,\|\boldsymbol{\psi}\|^2\,,$$

which, modulo the irrelevant constant $\|\mathbf{X}\|^2$, is the same as $\mathcal{C}_0$ of (3) with $\mathbf{X}$ replaced by $\mathbf{B}_u^T\mathbf{X}\mathbf{B}_v$, $\mathbf{u}$ by $\boldsymbol{\phi}$, and $\mathbf{v}$ by $\boldsymbol{\psi}$. Plain rank-one approximations $\boldsymbol{\phi}\boldsymbol{\psi}^T$ of $\tilde{\mathbf{X}}$ translate to regularized rank-one approximations $\mathbf{u}\mathbf{v}^T$ of $\mathbf{X}$ by $\mathbf{u} = \mathbf{B}_u\boldsymbol{\phi}$ and $\mathbf{v} = \mathbf{B}_v\boldsymbol{\psi}$. Applying plain SVD to the projected data matrix has formal similarity to constrained canonical correlation analysis (Takane et al., 2006), where subspace restrictions are used to incorporate constraints on both rows and columns of a data matrix in canonical correlation analysis.

The advantage of the basis expansion approach is its simplicity and less computational cost for large data sets. In general, however, basis expansion provides less flexible regularization than penalization. Computational cost of the penalization method can be reduced by truncating the full basis, as in pseudosplines (Hastie, 1996), or by applying penalization to the coefficients in a rich basis expansion, as

in penalized splines (Eilers and Marx, 1996; Ruppert et al., 2003). It is important to point out that our penalization framework allows powerful generalizations of the SVD through kernelization (Schölkopf and Smola, 2001), which the basis approach does not.

5 A connection with canonical correlation analysis

There exists a formal connection between two-way penalized SVDs and functional canonical correlation analysis (CCA) as introduced by Leurgans et al. (1993). The gist is that two-way penalized SVDs work the same way on data matrices as penalized CCA on sphered covariance matrices. To see this, we need to optimize the scale of the rank-one approximation $\mathbf{u}\mathbf{v}^T$:

$$\min_{s,t} \mathcal{C}(s\mathbf{u}, t\mathbf{v}) = \min_{r} \left(\|\mathbf{X}\|^2 - 2\,r\,\mathbf{u}^T\mathbf{X}\mathbf{v} + r^2\,\mathbf{u}^T(\mathbf{I}+\mathbf{\Omega}_u)\mathbf{u}\cdot\mathbf{v}^T(\mathbf{I}+\mathbf{\Omega}_v)\,\mathbf{v} \right) \quad (25)$$

$$= \|\mathbf{X}\|^2 - \frac{(\mathbf{u}^T\mathbf{X}\mathbf{v})^2}{\mathbf{u}^T(\mathbf{I}+\mathbf{\Omega}_u)\mathbf{u}\cdot\mathbf{v}^T(\mathbf{I}+\mathbf{\Omega}_v)\mathbf{v}}. \quad (26)$$

We call the last term a "bi-Rayleigh quotient" as it is a Rayleigh quotient (ratio of quadratics) in $\mathbf{u}$ conditional on $\mathbf{v}$, and vice versa:

$$\mathcal{R}(\mathbf{u},\mathbf{v}) = \frac{(\mathbf{u}^T\mathbf{X}\mathbf{v})^2}{\mathbf{u}^T(\mathbf{I}+\mathbf{\Omega}_u)\mathbf{u}\cdot\mathbf{v}^T(\mathbf{I}+\mathbf{\Omega}_v)\mathbf{v}}. \quad (27)$$

Specializing to $\mathbf{\Omega}_u = \mathbf{0}$ and $\mathbf{\Omega}_v = \mathbf{0}$, we obtain the unpenalized bi-Rayleigh quotient $\mathcal{R}_0(\mathbf{u},\mathbf{v}) = (\mathbf{u}^T\mathbf{X}\mathbf{v})^2/(\mathbf{u}^2\mathbf{v}^2)$ corresponding to $\mathcal{C}_0(\mathbf{u},\mathbf{v})$ in (3) above. Maximization of $\mathcal{R}(\mathbf{u},\mathbf{v})$ is equivalent to minimization of $\mathcal{C}(\mathbf{u},\mathbf{v})$ up to an undetermined slope factor. The stationary solutions $(\mathbf{u},\mathbf{v})$ of $\mathcal{R}(\mathbf{u},\mathbf{v})$ are pairs of singular vectors, and $\mathcal{R}(\mathbf{u},\mathbf{v})$ evaluated at singular vectors is the squared singular value (just as an ordinary Rayleigh quotient evaluated at an eigenvector is the eigenvalue).

The formal link to functional CCA is as follows: Given two variable blocks $\mathbf{X}$ and $\mathbf{Y}$ of sizes $n \times m_X$ and $n \times m_Y$, respectively, form the covariance matrices $\mathbf{C}_{X,X}$, $\mathbf{C}_{Y,Y}$ and $\mathbf{C}_{X,Y}$. Plain CCA is then obtained from the stationary solutions of the squared

correlation (= bi-Rayleigh quotient) $(\mathbf{u}^T\mathbf{C}_{X,Y}\mathbf{v})^2/(\mathbf{u}^T\mathbf{C}_{X,X}\mathbf{u}\cdot\mathbf{v}^T\mathbf{C}_{Y,Y}\mathbf{v})$, whose values at the stationary solutions are the squared canonical correlations. Penalized CCA (Leurgans et al. (1993), their equation (5)) is obtained from the stationary solutions of the penalized squared correlation (= another bi-Rayleigh quotient)

$$\frac{(\mathbf{u}^T\mathbf{C}_{X,Y}\mathbf{v})^2}{\mathbf{u}^T(\mathbf{C}_{X,X}+\mathbf{\Omega}_u)\mathbf{u}\cdot\mathbf{v}^T(\mathbf{C}_{Y,Y}+\mathbf{\Omega}_v)\mathbf{v}} \;=\; \frac{(\tilde{\mathbf{u}}^T\tilde{\mathbf{C}}_{X,Y}\tilde{\mathbf{v}})^2}{\tilde{\mathbf{u}}^T(\mathbf{I}+\tilde{\mathbf{\Omega}}_u)\tilde{\mathbf{u}}\cdot\tilde{\mathbf{v}}^T(\mathbf{I}+\tilde{\mathbf{\Omega}}_v)\tilde{\mathbf{v}}} \qquad (28)$$

where the right hand side derives from the sphering transformations $\tilde{\mathbf{u}} = \mathbf{C}_{X,X}^{1/2}\mathbf{u}$, $\tilde{\mathbf{v}} = \mathbf{C}_{Y,Y}^{1/2}\mathbf{v}$, $\tilde{\mathbf{C}}_{X,Y} = \mathbf{C}_{X,X}^{-1/2}\mathbf{C}_{X,Y}\mathbf{C}_{Y,Y}^{-1/2}$, $\tilde{\mathbf{\Omega}}_u = \mathbf{C}_{X,X}^{-1/2}\mathbf{\Omega}_u\mathbf{C}_{X,X}^{-1/2}$, $\tilde{\mathbf{\Omega}}_v = \mathbf{C}_{Y,Y}^{-1/2}\mathbf{\Omega}_v\mathbf{C}_{Y,Y}^{-1/2}$. The correspondence between (27) and the r.h.s. of (28) shows that two-way penalized SVDs and penalized CCA use the same formalism applied to different input matrices.

6 Data Examples

6.1 A Simulated Example

We illustrate regularized SVDs with simulated data sets generated from a model with two pairs of smooth components. These are specified as follows:

$$\begin{aligned} U_1^*(s) &= \sin(2\pi s)\,, & V_1^*(t) &= -3 + 8\exp(-4(t-0.25)^2)\,, \\ U_2^*(s) &= \sin(2\pi(s-0.25))\,, & V_2^*(t) &= -3 + 8\exp(-4(t-0.75)^2)\,. \end{aligned}$$

If s_i and t_j are each 201 equi-spaced points in $[0,1]$, the signal on the 201^2 grid is

$$x_{ij}^* \;=\; U_1^*(s_i)\,V_1^*(t_j) + U_2^*(s_i)\,V_2^*(t_j)\,.$$

Noisy data are generated by $x_{ij} = x_{ij}^* + e_{ij}$, where e_{ij} are independent Gaussian $\mathcal{N}(0,\sigma^2)$. The data matrix $\mathbf{X} = (x_{ij})$ was simulated 100 times. — The defining decomposition of x_{ij}^* is not in SVD form because the components are not orthogonal. To create a reasonable target for gauging the performance of the regularized SVD, we obtained a plain SVD of the noise-free matrix $\mathbf{X}^* = (x_{ij}^*)$, resulting in a decomposition

$$x_{ij}^* \;=\; d_1U_1(s_i)\,V_1(t_j) + d_2U_2(s_i)\,V_2(t_j)\,, \qquad (29)$$

where $\mathbf{u}_j = (U_j(s_1), \ldots, U_j(s_{201}))^T$, $\mathbf{v}_j = (V_j(t_1), \ldots, V_j(t_{201}))^T$, $\mathbf{u}_1^T\mathbf{u}_2 = \mathbf{v}_1^T\mathbf{v}_2 = 0$, $\|\mathbf{u}_j\|^2 = \|\mathbf{v}_j\|^2 = 1$. If successful, the regularized SVD will recover $\mathbf{u}_1$, $\mathbf{u}_2$, $\mathbf{v}_1$ and $\mathbf{v}_2$.

To recover the "true" components $\mathbf{u}_1$, $\mathbf{u}_2$, $\mathbf{v}_1$, $\mathbf{v}_2$ from (29), we apply to the simulated data matrices both plain and penalized SVDs, for which sums of squared second differences are used as the roughness penalty and the GCV criterion given in Section 3 is used to select the penalty parameters. Both methods produce approximately unbiased estimates (not shown here). However, the regularized SVD has much less variance as evident in Figure 2 which shows that regularized SVDs lead to uniformly smaller variance compared to plain SVDs, and the reduction in variance is quite substantial (55% on average). When examining individual simulations, the plain SVD yields quite noisy component estimates, while the regularized SVD does a good job at denoising and recovering smooth components without adding substantial bias; the scree plot, obtained for GCV-optimized bandwidths, suggests that there are two pairs of underlying components, while the GCV curves illustrate that the penalty parameter selection is quite sharp.

We also apply to the simulated data matrices several variants of SVD that yield smooth components. One such variant smoothes the components obtained from the plain SVD using the smoothing spline smoother with a GCV selected smoothing parameter. The other two variants are our penalized SVDs with restriction to one-way regularization. They are essentially Silverman (1996)'s penalized PCA procedures (see Huang et al., 2008). We also consider the SVD regularized with basis expansion/subspace restriction onto quadratic B-splines.

For each simulated data set, extracted components from various methods are compared with the corresponding "true" components, and the integrated squared errors (ISE) are calculated; the ratio of the ISE for a method and the ISE for the penalized SVD is then computed. The averages of ratios from 100 simulated data sets are reported in Table 1. From the results, the benefits of smoothing in general and of

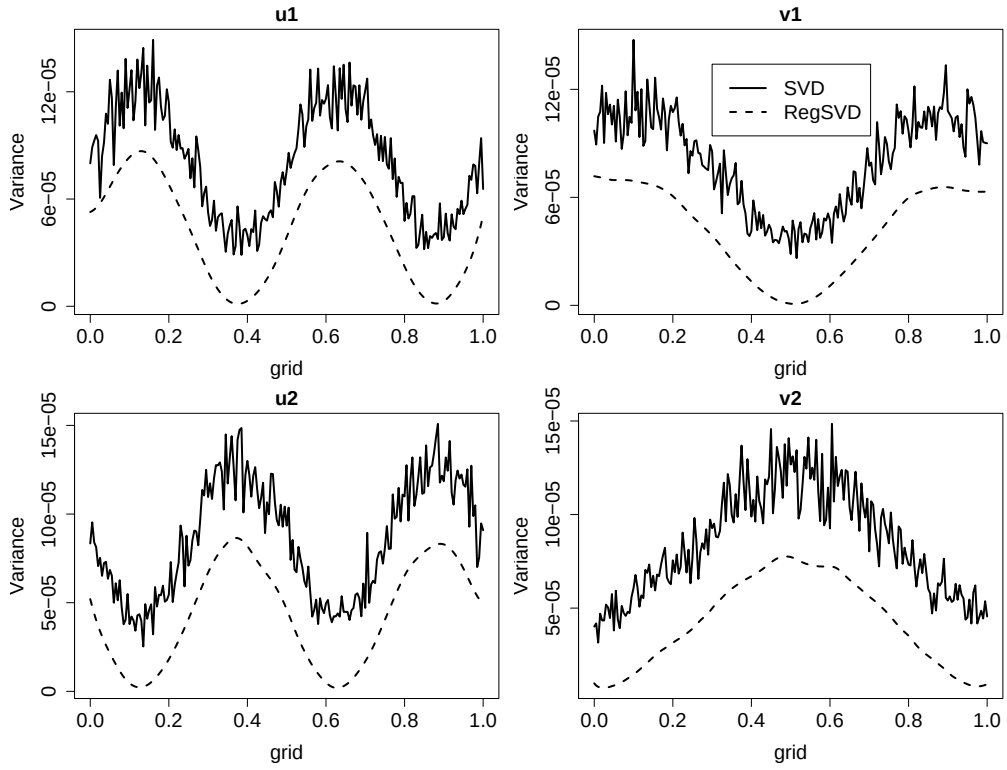


Figure 2: Pointwise variance comparison of components extracted with plain and penalized SVDs. The singular vectors are standardized to norm 1. Noise level $\sigma = 3$.

our approach become quite clear: none of the values is below one, that is, none of the other approaches beats ours. The naive method of smoothing the noisy components from plain SVD is inferior to our more principled regularization approach, especially when the noise level is high. When only $\mathbf{u}$ (or $\mathbf{v}$) components are regularized, the $\mathbf{v}$ (or $\mathbf{u}$) components are badly recovered and even with smoothing, the recovery of the $\mathbf{u}$ (or $\mathbf{v}$) components deteriorates. The SVD regularized with basis expansion gives results similar but slightly inferior to the SVD regularized with penalization.

In the two real data examples below we shall focus on penalized SVDs. We form the penalty matrices as in smoothing splines (Green and Silverman, 1994) and use the GCV criteria in Section 3.1 to select the penalty parameters.

Methods	noise level σ	$\mathbf{u}_1$	$\mathbf{u}_2$	$\mathbf{v}_1$	$\mathbf{v}_2$
SVD	3	7.48 (0.89)	7.69 (0.81)	12.31 (2.76)	9.50 (1.18)
	6	8.08 (0.94)	9.26 (1.04)	15.33 (3.07)	11.94 (1.57)
sSVD	3	1.09 (0.03)	1.07 (0.03)	1.06 (0.04)	1.10 (0.04)
	6	1.38 (0.14)	1.38 (0.14)	1.64 (0.28)	1.56 (0.22)
uSVD	3	1.07 (0.03)	1.06 (0.02)	12.17 (2.70)	9.39 (1.17)
	6	1.22 (0.06)	1.19 (0.05)	14.49 (2.84)	11.29 (1.45)
vSVD	3	7.37 (0.88)	7.57 (0.80)	1.03 (0.03)	1.02 (0.03)
	6	7.65 (0.89)	8.74 (1.00)	1.30 (0.16)	1.24 (0.12)
rSVD-basis	3	1.05 (0.01)	1.08 (0.02)	1.10 (0.03)	1.10 (0.02)
	6	1.08 (0.05)	1.14 (0.05)	1.22 (0.09)	1.19 (0.06)

Table 1: Comparison of several methods with penalized SVD. Numbers reported are the means (SEs) of the ratios between the ISEs for a specified method and the ISEs for the penalized SVD, based on 100 simulation runs. The methods are plain SVD (SVD), plain SVD followed by smoothing (sSVD), penalized SVD that regularized $\mathbf{u}$ only (uSVD), penalized SVD that regularized $\mathbf{v}$ only (vSVD), and regularized SVD using quadratic spline basis expansion (rSVD-basis).

6.2 Example: US Mortality Rate Data

There is a literature on mortality forecasting based on the Lee-Carter method and its extensions (Lee and Carter, 1992; Hyndmann and Booth, 2008), where the SVD is combined with time series modeling. Now we use our regularized SVD to model mortality rates from a smoothing perspective. While our method is not intended for prediction, it does help reveal some interesting phenomena of mortality transition.

We use the US mortality rate data from the Berkeley Human Mortality Database (http://www.mortality.org/). These data, previously analyzed in Yang et al. (2004), contain mortality rates in the United States for ages 0 to 95 from 1959 to 1999. We will focus on female mortality rates. The data matrix $\mathbf{X}$ is of size 41×96, each row

corresponding to a one-year period and each column to an age group. Prior to their analysis, Yang et al. (2004) aggregated the mortality rates into 5-year *Age* groups and 5-year *Period* groups. We will replace such data aggregation with the smoothing implicit in regularized SVDs and hence work with the un-aggregated data.

We assume the observed data matrix $\mathbf{X}$ to be a discretization of an underlying two-way smooth function $X(\text{Period}, \text{Age})$. The regularized SVD fits the following model for explaining the mortality rate in terms of Period and Age:

$$X(\text{Period}, \text{Age}) \;=\; d_1 U_1(\text{Period})\, V_1(\text{Age}) + \ldots + d_q U_q(\text{Period})\, V_q(\text{Age}) + \epsilon(\text{Period}, \text{Age})\,,$$

where $U_i(\cdot)$ and $V_j(\cdot)$ are smooth functions of Period and Age, respectively. Note that the fitted $U_1(\cdot)$ and $V_1(\cdot)$ should not be interpreted as mean functions, and $U_k(\cdot)$ and $V_k(\cdot)$ not as the $(k-1)$th principal components. They are just the components in the best fitting model with product terms.

We apply plain and regularized SVDs to the data for the age groups up to 95. For the plain SVD, the first pair, $(\mathbf{u}_1, \mathbf{v}_1)$, explains about 99.87% of the total energy, while the second pair $(\mathbf{u}_2, \mathbf{v}_2)$ explains 83.11% of the remaining energy. Panel (a) of Figure 3 shows the proportion of remaining energy explained by components $k \leq 10$. We decided to use the first two pairs of components to summarize the data because they clearly separate from the rest of pairs. These components are plotted in panels (c)-(f) of the same figure, with a zoom of the plot of V_1 in panel (b) to show greater detail for Age ≤ 15.

The curve V_1 shows the well-known pattern of mortality age curves (Wilmoth, 1990): declining sharply between age 0 and age 2, slowly dropping to a minimum around 12, rising to a local mild peak in the late teens, leveling off for the next decade before increasing exponentially after age 30. The curve U_1 exhibits the smooth *average* time trend across periods, suggesting the following period-mortality pattern: a persistent decline between 1959 and 1982, a flat pattern afterwards until 1987, then a continuous drop until 1993 and finally a slight increase.

The second component focuses mainly on the early and late ages where it corrects for patterns that the first component was unable to cover. The curve V_2 takes a positive value at low ages and between ages 70 and 90. The curve U_2 shows a gradual decrease from positive to negative in the time period under consideration, capturing mostly a contrast between the 1960s and the 1990s. The combined message is that, for ages under two and between 70 and 90, the mortality rate is higher in the 1960s and lower in the 1990s than what can be explained by the one component model.

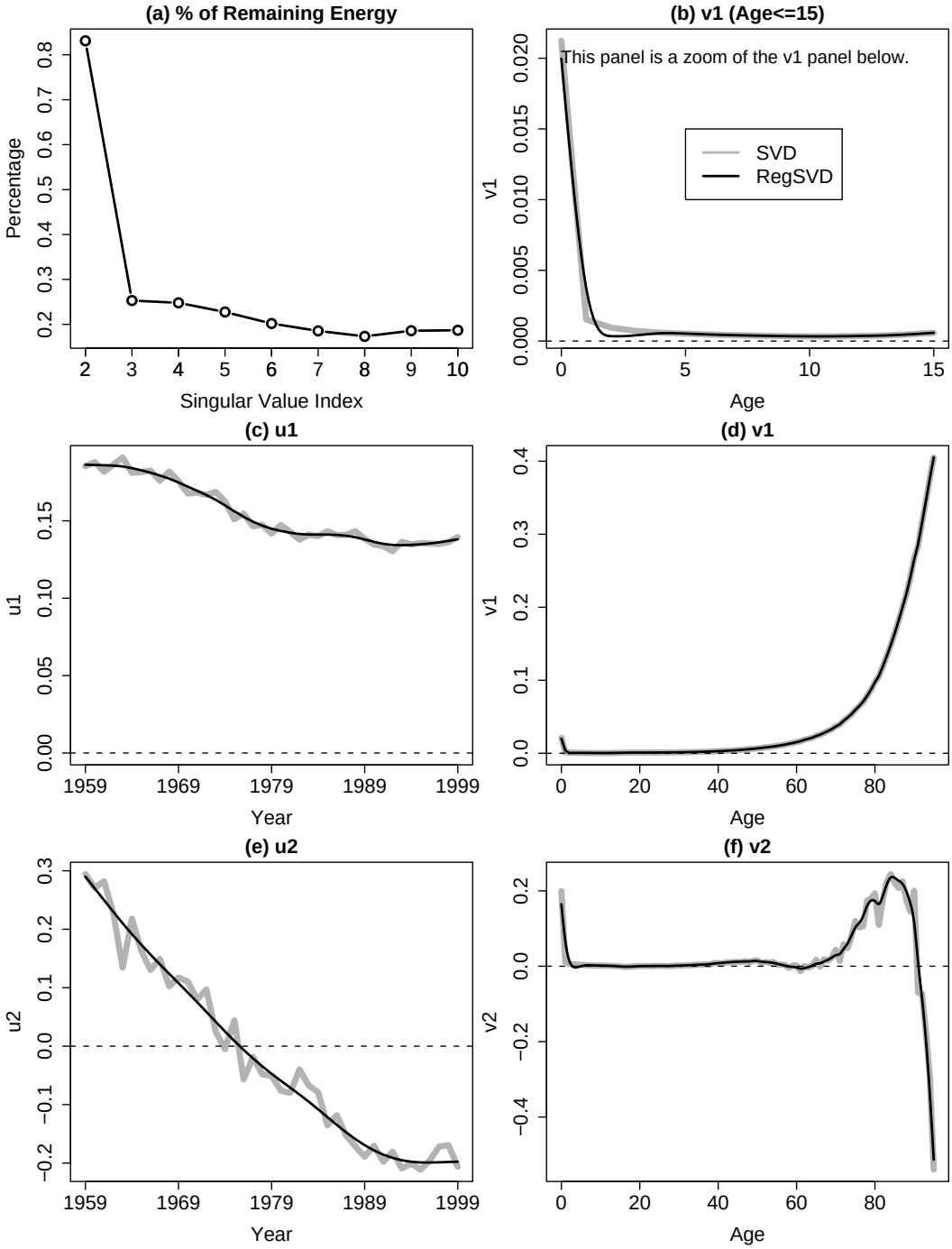
6.3 Example: Israeli Call Center Customer Patience Data

We apply the regularized SVD to the Israeli call center data analyzed in Brown et al. (2005). Call centers have become a primary communication channel between companies and their customers in modern business. Various aspects of call center operations were first subjected to a thorough statistical analysis by Brown et al. (2005) who analyzed for example arrival processes, service durations, and customer patience.

To illustrate regularized SVDs, we study customer patience as a function of both time of day and waiting time. Customer patience is an issue (as we all know) because customers wait in a virtual telephone queue before receiving service. Eventually, a customer either gets served or hangs up if patience runs out. Brown et al. (2005) emphasize the importance of understanding customer patience for efficient system design and call routing. They propose *the time a customer is willing to wait before hanging up* as an ancillary measure for patience. Denote this time by W; it is observed only if the customer does in fact hang up; if the customer is served, W is right-censored.

The data we analyze focus on all the agent-seeking customer calls that got connected to the center between 07:00 and midnight during every weekday in November and December of 1999. In particular, for each customer, the data record the arrival time of the call, the waiting time, and whether the customer gets served or hangs up.

Figure 3: US Female Mortality Rate: Panel (a) shows a renormalized scree plot after removal of the dominant first component. Panel (b) shows a zoom of the bottom left corner of panel (d).



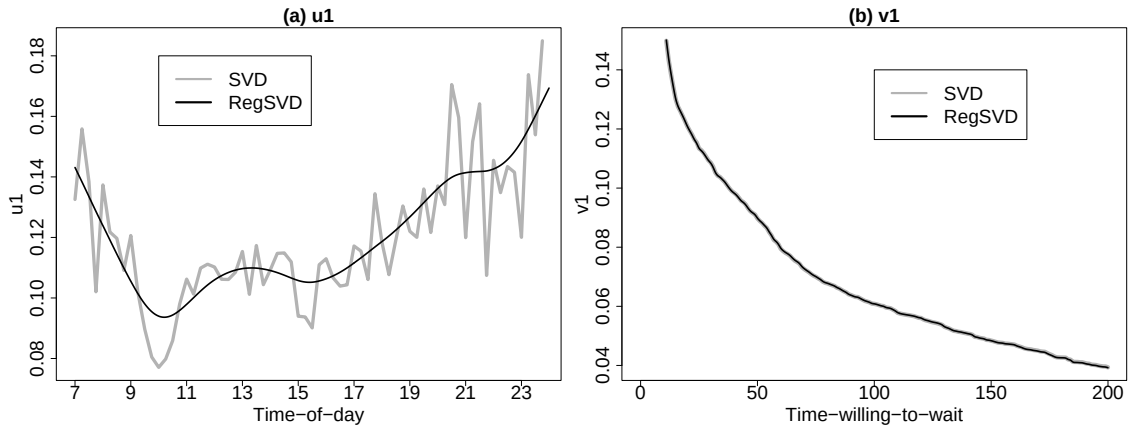


Figure 4: Israeli Customer Patience: Comparison of the First Components

Following common practice, we first group all the calls into 68 quarter hours from 07:00 to midnight. For each interval, we apply the Kaplan-Meier estimator (Kaplan and Meier, 1958) to obtain the survival function of time-willing-to-wait W, with which we then calculate the log-odds function of patience $\log\{P(W > w)/P(W \leq w)\}$. One reason for considering log-odds is that they are interval scale (use the whole real line), which renders them more appropriate for an SVD-based analysis. The final data matrix $\mathbf{X}$ consists — for each quarter hour interval — of the evaluations of the log-odds function at the seconds 11, 12, ..., 200 for the waiting times. Hence the size of $\mathbf{X}$ is 68×190, where the rows are indexed by 15-minute time-of-day intervals, and the columns are indexed by waiting times in seconds for all seconds from 11 to 200.

The regularized SVD yields the following model of the log-odds as a function of time-of-day t and time-willing-to-wait w,

$$X(t,w) \;=\; d_1 U_1(t)\, V_1(w) \;+\; \ldots \;+\; d_q U_q(t)\, V_q(w) \;+\; \epsilon(t,w)\,, \tag{30}$$

where $U_i(\cdot)$ and $V_i(\cdot)$ are smooth in time-of-day and time-willing-to-wait, respectively.

Figure 4 compares the first pair of components between plain and regularized SVDs. In Panel (a) the regularized singular curve reveals an interesting double-dip pattern of log-odds as a function of time-of-day. The function decreases from the

early morning, reaches the first valley around 10:00, increases afterwards until 13:00, decreases again until the second valley around 15:00, before increasing until midnight. According to this plot, customers are the least patient around 10:00 and 15:00, which happen to be the peak hours with the most call arrivals (Brown et al., 2005). This suggests that customers are more likely to hang up while there are more customers. This observation seems intuitive and complements the findings in Brown et al. It certainly deserves further investigation because of its obvious interest to call centers.

For the plain SVD, the first pair, $(\mathbf{u}_1, \mathbf{v}_1)$, explains about 98.93% of the total energy, which suggests that the first pair summarizes the dominating mode of variation in the data. We stop at the first pair because a plot similar to Figure 3(a) does not separate the second pair from the rest in explaining the remaining energy (not shown). Model (30) with one SVD component is essentially a proportional log-odds model, where $V_1(w)$ captures the baseline pattern and $d_1 U_1(t)$ provides the time-of-day specific scale adjustment. This model suggests that customers seem to have the same aggregate behavior in terms of time-willing-to-wait at different times of day.

Appendix: Proofs of Theorem 1 and Lemma 1

Proof of Theorem 1: The penalty $\mathcal{P}(\mathbf{u}, \mathbf{v})$ is assumed to be a quadratic in both arguments, but the quadratics may depend on the other argument, hence $\mathcal{P}(\mathbf{u}, \mathbf{v}) = \mathbf{u}^T \mathbf{A}(\mathbf{v})\mathbf{u} = \mathbf{v}^T \mathbf{B}(\mathbf{u})\mathbf{v}$, where $\mathbf{A}(\mathbf{v})$ and $\mathbf{B}(\mathbf{u})$ are symmetric and of suitable sizes. The criterion can therefore be written in two ways:

$$\begin{aligned} \mathcal{C}(\mathbf{u}, \mathbf{v}) &= \|\mathbf{X}\|^2 - 2\mathbf{u}^T \mathbf{X}\mathbf{v} + \mathbf{u}^T(\|\mathbf{v}\|^2 \mathbf{I} + \mathbf{A}(\mathbf{v}))\mathbf{u} \qquad (31) \\ &= \|\mathbf{X}\|^2 - 2\mathbf{u}^T \mathbf{X}\mathbf{v} + \mathbf{v}^T(\|\mathbf{u}\|^2 \mathbf{I} + \mathbf{B}(\mathbf{u}))\mathbf{v}\,. \qquad (32) \end{aligned}$$

The stationarity condition for $\mathbf{u}$ given a fixed $\mathbf{v}$ is

$$\frac{\partial}{\partial \mathbf{u}} \mathcal{C}(\mathbf{u}, \mathbf{v}) \;=\; -2(\,\mathbf{X}\mathbf{v} + (\|\mathbf{v}\|^2 \mathbf{I} + \mathbf{A}(\mathbf{v}))\mathbf{u}\,) \;=\; \mathbf{0}\,.$$

The stationary solution $\mathbf{u}$ is therefore $\mathbf{u} = (\|\mathbf{v}\|^2\mathbf{I} + \mathbf{A}(\mathbf{v}))^{-1}\mathbf{X}\mathbf{v}$. Next we use the argmin assumption which implies that the stationary solution is of the form $\mathbf{S}_u\mathbf{X}\mathbf{v}/g(\mathbf{v})$, where $g(\mathbf{v})$ is a scalar function and the reciprocal is chosen for subsequent convenience. It follows $(\|\mathbf{v}\|^2\mathbf{I}+\mathbf{A}(\mathbf{v}))^{-1} = \mathbf{S}_u/g(\mathbf{v})$, and hence $\|\mathbf{v}\|^2\mathbf{I}+\mathbf{A}(\mathbf{v}) = (\mathbf{I}+\mathbf{\Omega}_u)g(\mathbf{v})$. Substituting in (31) shows

$$\mathcal{C}(\mathbf{u},\mathbf{v}) \;=\; \|\mathbf{X}\|^2 - 2\mathbf{u}^T\mathbf{X}\mathbf{v} + \mathbf{u}^T(\mathbf{I}+\mathbf{\Omega}_u)\mathbf{u}\cdot g(\mathbf{v})\,.$$

The dual result with the roles of $\mathbf{u}$ and $\mathbf{v}$ exchanged is

$$\mathcal{C}(\mathbf{u},\mathbf{v}) \;=\; \|\mathbf{X}\|^2 - 2\mathbf{u}^T\mathbf{X}\mathbf{v} + f(\mathbf{u})\cdot\mathbf{v}^T(\mathbf{I}+\mathbf{\Omega}_v)\mathbf{v}\,.$$

Equating the last summand of each we find that

$$\frac{f(\mathbf{u})}{\mathbf{u}^T(\mathbf{I}+\mathbf{\Omega}_u)\mathbf{u}} \;=\; \frac{g(\mathbf{v})}{\mathbf{v}^T(\mathbf{I}+\mathbf{\Omega}_v)\mathbf{v}} \;=\; \lambda$$

must be constant. Therefore,

$$\mathcal{C}(\mathbf{u},\mathbf{v}) \;=\; \|\mathbf{X}\|^2 - 2\mathbf{u}^T\mathbf{X}\mathbf{v} + \lambda\,\mathbf{u}^T(\mathbf{I}+\mathbf{\Omega}_u)\mathbf{u}\cdot\mathbf{v}^T(\mathbf{I}+\mathbf{\Omega}_v)\mathbf{v} \tag{33}$$

For $\mathbf{\Omega}_u = \mathbf{0}$ and $\mathbf{\Omega}_v = \mathbf{0}$, this specializes to $\mathcal{C}(\mathbf{u},\mathbf{v}) \;=\; \|\mathbf{X}\|^2 - 2\mathbf{u}^T\mathbf{X}\mathbf{v} + \lambda\,\|\mathbf{u}\|^2\cdot\|\mathbf{v}\|^2$, which, in view of (3), reduces to $\mathcal{C}_0(\mathbf{u},\mathbf{v}) = \|\mathbf{X}-\mathbf{u}\mathbf{v}^T\|^2$ only for $\lambda = 1$. □

Proof of Lemma 1: Deleting one column of $\mathbf{X}$ corresponds to deleting a block of size n from $\bar{\mathbf{y}}$ in the ridge regression (22). Partition the hat matrix $\mathbf{H}$ into $m \times m$ equal-sized blocks where each block corresponds to a column of $\mathbf{X}$. Note that $\hat{\mathbf{v}}^{(-j)}$ also solves the ridge regression (22) when the j-th block of $\bar{\mathbf{y}}$ is replaced by $\mathbf{u}\hat{v}_j^{(-j)}$. The j-th block of the fitted equation $\hat{\bar{\mathbf{y}}} = \mathbf{H}\bar{\mathbf{y}}$ of this latter ridge regression reads as

$$\mathbf{u}\hat{v}_j^{(-j)} = \sum_{k\neq j}\mathbf{H}_{jk}\mathbf{x}_k + \mathbf{H}_{jj}\{\mathbf{u}\hat{v}_j^{(-j)}\}.$$

Subtracting $\mathbf{x}_j$ on both sides of the above identity and observing that $\sum_k \mathbf{H}_{jk}\mathbf{x}_k = \mathbf{u}\hat{\mathbf{v}}_j$, we obtain

$$\begin{aligned}\mathbf{u}\hat{v}_j^{(-j)} - \mathbf{x}_j &= \sum_{k}\mathbf{H}_{jk}\mathbf{x}_k - \mathbf{x}_j + \mathbf{H}_{jj}\{\mathbf{u}\hat{v}_j^{(-j)} - \mathbf{x}_j\}\\ &= \mathbf{u}\hat{v}_j - \mathbf{x}_j + \mathbf{H}_{jj}\{\mathbf{u}\hat{v}_j^{(-j)} - \mathbf{x}_j\}.\end{aligned}$$

Therefore, the cross-validated residual for deleting the j-th column of $\mathbf{X}$ is

$$\mathbf{u}\hat{v}_j^{(-j)} - \mathbf{x}_j = (\mathbf{I} - \mathbf{H}_{jj})^{-1}(\mathbf{u}\hat{v}_j - \mathbf{x}_j),$$

where

$$\mathbf{H}_{jj} = \frac{(\mathbf{S}_v)_{jj}}{\mathbf{u}^T(\mathbf{I} + \mathbf{\Omega}_u)\mathbf{u}}\mathbf{u}\mathbf{u}^T \triangleq \gamma_j \mathbf{u}\mathbf{u}^T$$

and $\gamma_j\|\mathbf{u}\|^2 = (\mathbf{S}_v)_{jj}/(1 + \alpha_u\mathcal{R}_u)$ with $\mathcal{R}_u = \mathbf{u}^T\mathbf{\Omega}_u\mathbf{u}/\mathbf{u}^T\mathbf{u}$. Denote $\mathbf{w} = \mathbf{u}\hat{v}_j - \mathbf{x}_j$. Its squared norm is

$$\begin{aligned}\|\mathbf{w}\|^2 &= \mathbf{x}_j^T\mathbf{x}_j - 2\mathbf{x}_j^T\mathbf{u}\hat{v}_j + \|\mathbf{u}\|^2\hat{v}_j^2 \\ &= \mathbf{x}_j^T\mathbf{x}_j - \frac{(\mathbf{x}_j^T\mathbf{u})^2}{\|\mathbf{u}\|^2} + \left(\|\mathbf{u}\|\hat{v}_j - \frac{\mathbf{u}^T\mathbf{x}_j}{\|\mathbf{u}\|}\right)^2.\end{aligned} \tag{34}$$

Since $\mathbf{u}^T\mathbf{w} = \|\mathbf{u}\|^2(\hat{v}_j - \mathbf{u}^T\mathbf{x}_j/\|\mathbf{u}\|^2)$, we have that

$$\frac{(\mathbf{u}^T\mathbf{w})^2}{\|\mathbf{u}\|^2} = \left(\|\mathbf{u}\|\hat{v}_j - \frac{\mathbf{u}^T\mathbf{x}_j}{\|\mathbf{u}\|}\right)^2. \tag{35}$$

Using the identity

$$(\mathbf{I} - \gamma_j\mathbf{u}\mathbf{u}^T)^{-1} = \mathbf{I} + \frac{\gamma_j}{1 - \gamma_j\|\mathbf{u}\|^2}\mathbf{u}\mathbf{u}^T,$$

we can write the cross-validated residual $\mathbf{u}\hat{v}_j^{(-j)} - \mathbf{x}_j$ as

$$\begin{aligned}(\mathbf{I} - \mathbf{H}_{jj})^{-1}(\mathbf{u}\hat{v}_j - \mathbf{x}_j) &= \left(\mathbf{I} + \frac{\gamma_j}{1 - \gamma_j\|\mathbf{u}\|^2}\mathbf{u}\mathbf{u}^T\right)\mathbf{w} \\ &= \mathbf{w} + \frac{\gamma_j}{1 - \gamma_j\|\mathbf{u}\|^2}(\mathbf{u}^T\mathbf{w})\mathbf{u}.\end{aligned}$$

Thus the squared norm of $\mathbf{u}\hat{v}_j^{(-j)} - \mathbf{x}_j$ is

$$\begin{aligned}&\|\mathbf{w}\|^2 + \frac{2\gamma_j}{1 - \gamma_j\|\mathbf{u}\|^2}(\mathbf{u}^T\mathbf{w})^2 + \frac{\gamma_j^2}{(1 - \gamma_j\|\mathbf{u}\|^2)^2}(\mathbf{u}^T\mathbf{w})^2\|\mathbf{u}\|^2 \\ &= \|\mathbf{w}\|^2 + \frac{(\mathbf{u}^T\mathbf{w})^2}{\|\mathbf{u}\|^2}\left\{\frac{2\gamma_j\|\mathbf{u}\|^2}{(1 - \gamma_j\|\mathbf{u}\|^2)} + \frac{\gamma_j^2\|\mathbf{u}\|^4}{(1 - \gamma_j\|\mathbf{u}\|^2)^2}\right\} \\ &= \|\mathbf{w}\|^2 + \frac{(\mathbf{u}^T\mathbf{w})^2}{\|\mathbf{u}\|^2}\left\{\frac{1}{(1 - \gamma_j\|\mathbf{u}\|^2)^2} - 1\right\}.\end{aligned}$$

Combining this result with (34) and (35) we obtain (23). □

Supplemental Materials

Several Flawed Approaches to Penalized SVDs: In this note, we show that several "natural" approaches to penalized SVDs do not work and explain why so. (pdf file)

References

Brown, L. D., Gans, N., Mandelbaum, A., Sakov, A., Shen, H., Zeltyn, S., and Zhao, L. (2005), "Statistical analysis of a telephone call center: a queueing-science perspective," *Journal of the American Statistical Association*, 100, 36–50.

Crellin, N. J. (1996), "Modeling Image Sequences, with Particular Application to FMRI Data," Ph.D. thesis, Department of Statistics, Stanford University.

Eilers, P. and Marx, B. (1996), "Flexible smoothing with B-splines and penalties (with discussion)," *Statistical Science*, 11, 89–121.

Golub, G. H. and van Loan, C. F. (1996), *Matrix Computation*, The Johns Hopkins University Press, 3rd ed.

Green, P. and Silverman, B. (1994), *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach*, Chapman & Hall.

Hastie, T. (1996), "Pseudosplines," *Journal of the Royal Statistical Society, Series B*, 58, 379–396.

Hastie, T. and Tibshirani, R. (1990), *Generalized Additive Models*, London, UK: Chapman and Hall.

Huang, J., Shen, H., and Buja, A. (2008), "Functional principal components analysis via penalized rank one approximation," *Electronic Journal of Statistics*, 2, 678–695.

Hyndmann, R. J. and Booth, H. (2008), "Stochastic population forecasts using functional data models for mortality, fertility and migration," *International Journal of Forecasting*, 24, 323–342.

Kaplan, E. L. and Meier, P. (1958), "Nonparametric estimation from incomplete observations," *Journal of the American Statistical Association*, 53, 457–481.

Kimeldorf, G. S. and Wahba, G. (1971), "Some results on Tchebycheffian spline functions," *Journal of Mathematical Analysis and Applications*, 33, 82–95.

Lee, R. D. and Carter, L. R. (1992), "Modelling and forecasting U.S. mortality," *Journal of the American Statistical Association*, 87, 659–675.

Leurgans, S., Moyeed, R. A., and Silverman, B. W. (1993), "Canonical Correlation Analysis when the Data are Curves," *Journal of the Royal Statistical Society, Series B*, 55, 725–740.

Mandel, J. (1971), "A New Analysis of Variance Model for Non-Additive Data," *Technometrics*, 13, 1–18.

Ramsay, J. O. and Silverman, B. W. (2002), *Applied Functional Data Analysis*, New York, NY: Springer-Verlag.

— (2005), *Functional Data Analysis*, New York, NY: Springer-Verlag, 2nd ed.

Rao, C. R. (1958), "Some statistical methods for comparison of growth curves," *Biometrics*, 14, 1–17.

— (1987), "Prediction in growth curve models (with discussion)," *Statistical Science*, 2, 434–471.

Rice, J. A. and Silverman, B. W. (1991), "Estimating the mean and Covariance structure nonparametrically when the data are curves," *Journal of the Royal Statistical Society, Series B*, 53, 233–243.

Ruppert, D., Wand, M. P., and Carroll, R. J. (2003), *Semiparametric regression*, Cambridge University Press.

Schölkopf, B. and Smola, A. J. (2001), *Learning with Kernels*, Cambridge, MA: MIT Press.

Silverman, B. W. (1996), "Smoothed functional principal components analysis by choice of norm," *The Annals of Statistics*, 24, 1–24.

Takane, Y., Yanai, H., and Hwang, H. (2006), "An improved method for generalized constrained canonical correlation analysis," *Computational Statistics and Data Analysis*, 50, 221–241.

Wahba, G. (1990), *Spline Models for Observational Data*, CBMS-NSF Regional Conference Series in Applied Mathematics, SIAM.

Williams, E. J. (1952), "The Interpretation of Interactions in Factorial Experiments," *Biometrika*, 39, 65–81.

Wilmoth, J. R. (1990), "Variation in vital rates by age, period, and cohort," *Sociological Methodology*, 20, 295–335.

Yang, Y., Fu, W. J., and Land, K. C. (2004), "A methodological comparison of age-period-cohort models: The intrinsic estimator and conventional generalized linear models," *Sociological Methodology*, 34, 75–110.